A New Framework for Underdetermined Speech Extraction Using Mixture of Beamformers

A New Framework for Underdetermined Speech Extraction Using Mixture of Beamformers

Mohammad A. Dmour *, *Student Member, IEEE,* and Mike Davies, *Member, IEEE*

Abstract

This paper describes frequency-domain non-linear mixture of beamformers that can extract a speech source from a known direction when there are fewer microphones than sources (the underdetermined case). Our approach models the data in each frequency bin via Gaussian mixture distributions, which can be learned using the expectation maximization algorithm. The model learning is performed using the observed mixture signals only, and no prior training is required. Non-linear beamformers are then developed based on this model. The proposed estimators are a non-linear weighted sum of linear minimum mean square error or minimum variance distortionless response beamformers. The resulting non-linear beamformers do not need to know or estimate the number of sources, and can be applied to microphone arrays with two or more microphones. We test and evaluate the described methods on underdetermined speech mixtures.

I. Introduction

Most audio signals result from the mixing of several sound sources. In many applications, there is a need to separate the multiple sources or extract a source of interest while reducing undesired interfering signals and noise. The estimated signals may then be either directly listened to or further processed, giving rise to a wide range of applications such as hearing aids, human-computer interaction, surveillance, and hands-free telephony.

Mohammad A. Dmour would like to acknowledge the support of the Wolfson Microelectronics Scholarship. Mike Davies acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the HeriotWatt University as a component part of the Edinburgh Research Partnership.

The authors are with IDCOM and the Joint Research Institute for Signal and Image Processing, Edinburgh University, Edinburgh EH9 3JL, U.K. (e-mail: M.Dmour@ed.ac.uk; mike.davies@ed.ac.uk)

There has been a lot of research on the speech enhancement problem, where the focus is on attenuating the background noise. Speech denoising algorithms are well established and have been used for many years [1], [2]. The extension of the speech enhancement problem to deal with mixtures of speech sources is a topic of intense research. Source mixing can occur in a wide variety of situations under different environments. The difficulty of source extraction depends on the way the source signals are mixed within the environment and on the a priori knowledge of the sources, microphones, and mixing parameters. Blind methods do not assume a priori knowledge of sources, microphones, or mixing parameters. By contrast, informed methods exploit some a priori information about the sources and microphones (for example, the location of a desired source). In general, the problem is more difficult when the reverberation time (RT) of the acoustic environment is large, and in the underdetermined case (fewer microphones than sources).

In general, observations are obtained at the output of a set of microphones, each receiving different combinations of the source signals. The use of microphone arrays gives one the opportunity to exploit the fact that the desired source and the interfering sources originate at different points in space. Suppose that M simultaneously active source signals are mixed and observed at N microphones. The signal recorded at the j^{th} microphone at time t can be modeled as:

$$x_j(t) = \sum_{i=1}^{M} \sum_{p=0}^{P-1} a_{ji}(p) s_i(t-p) \tag{1}$$

where a_{ji} represents the impulse response of the acoustic path from source i to microphone j and P is the length of the impulse response between each source-microphone pair. A mixture is termed a determined mixture when the number of microphones is equal to the number of sources, overdetermined when the number of microphones is larger than the number of sources, and underdetermined when it is smaller. In certain applications, source separation methods are used to estimate separated signals $\hat{s}_1, ..., \hat{s}_M$, which correspond to each of the original source signals $s_1, ..., s_M$. In many practical applications, however, prior information about a desired source, such as source location or identity, might be available and exploited to extract only one source of interest.

Various methods have been proposed for solving the speech separation problem. One approach is to use statistical modeling of source signals. Independent component analysis (ICA) is one of the major statistical tools used. In ICA, separation is performed on the assumption that the source signals are statistically independent, and does not require information on microphone array configuration or the direction of arrival (DOA) of the source signals to be available. To perform source separation, we process the mixture channels by a set of linear time-invariant demixing filters. ICA implicitly estimates the source directions

by maximizing the independence of the sources, and acts as an adaptive null beamformer that reduces the undesired sources. However, some aspects limit the application of ICA to real-world environments. Most ICA methods assume the number of sources is given a priori. In general, classical ICA techniques cannot perform source separation in the underdetermined mixtures case. For some excellent reviews of convolutive ICA methods for speech separation, see [3], [4]. Another strategy suited to speech mixtures is to incorporate speaker-independent models that can be learned from large speech datasets. For example, in [5], [6], Gaussian mixture models (GMMs) are used to model the sources, and the parameters of the mixing channel and noise were inferred using variational expectation maximization (EM) techniques [7]. The sources are then estimated using a minimum mean square error (MMSE) estimator. However, this method requires the number of sources to be known, and it cannot perform source separation in the underdetermined mixtures case.

A popular approach to speech extraction is to use adaptive beamforming techniques. With adaptive beamforming, the microphone array is also used to form a spatial filter which can extract a signal from a specific direction and reduce signals from other directions. For example, in minimum-variance distortionless response (MVDR) beamforming, the beamformer response is constrained so that signals from the direction of interest are passed with no distortion, while it suppresses noise and interference. In [8], [9], beamforming weights were calculated using time-domain recursive algorithms. It was shown recently in [10] that a frequency-domain MVDR (FMV) beamformer which performs sample matrix inversion using statistics estimated from a short sample support gives better performance than time-domain recursive algorithms in non-stationary acoustic environments. Unlike the ICA approach, adaptive beamforming requires information about the microphone array configuration and the sources (for example, the direction of the desired source). However, adaptive beamforming techniques can attenuate spatially spread and reverberant interferers, and there is no need to determine their number. In general, linear adaptive beamforming can attain excellent separation performance in determined or over-determined time-invariant mixtures. However, in underdetermined mixtures, perfect attenuation of all interferers becomes impossible and only partial interference attenuation is possible. This, in turn, leads to performance degradation.

In the underdetermined mixtures case, the assumption of spatial diversity alone is insufficient to perform source separation/extraction, thereby necessitating additional assumptions. The assumption that the sources have a sparse representation in a given basis is an increasingly popular addition. Sparseness of a signal means that only a few instances have a value significantly different from zero. A sparse representation of a speech signal can be achieved by a short term Fourier transform (STFT). One popular

approach to sparsity-based separation is time-frequency masking [11]–[17]. This approach is a special case of non-linear time-varying filtering that estimates the desired source s_i from a mixture signal x_j by:

$$\hat{s}_i(n, f) = M_i(n, f)x_j(n, f) \tag{2}$$

where $s_i(n, f)$ and $x_j(n, f)$ are the STFT coefficients of $s_i(t)$ and $x_j(t)$ respectively in the time frame n and frequency bin f, and M_i is a time-frequency mask containing positive gains which must be adapted to extract the desired source s_i from the observed mixture. A popular method to estimate the time-frequency masks using only two microphones is the degenerate unmixing estimation technique (DUET) [11], [12]. It is assumed that the time-frequency representation of speech signals are approximately disjoint (i.e., sources do not overlap too much):

$$s_i(n, f)s_j(n, f) \simeq 0, \qquad \forall i \neq j, \forall f \tag{3}$$

This assumption is not fully met in practice. In DUET, the source directions and the active source indices are alternately optimized by partitioning the mixture STFT coefficients based on the inter-channel level/phase difference (ILD/IPD). DUET is capable of performing separation of two or more sources using just two channels, and without significant computational complexity. However, this method suffers from so-called musical noise or burbling artifacts due to binary masking of time-frequency points where the sources overlap. An extension to the DUET algorithm for more than two microphones in was proposed in in [18], [19]. This method, Multiple sENsor dUET (MENUET), can be applied to non-linear microphone arrangements with 2- or 3-dimensional arrays. In [13]–[17], probabilistic models are used to model the IPD/ILD, and after estimating its parameters with an EM algorithm [20], soft masks can be derived. All of these methods require the number of sources to be given a priori, and it is difficult to expand these methods to more than two microphones. Furthermore, separation methods based on time-frequency masking suffer from the fact that clustering becomes difficult in reverberation, as ILD/IPD resulting from each sound source then tend to spread and overlap.

When only one microphone is available, source separation/extraction becomes significantly more challenging, as spatial cues are absent in this case. In this situation, the assumptions of independence and time-frequency sparsity becomes insufficient, and more advanced source models relying on spectro-temporal models are needed. Different strategies have been employed using these models [21]–[23]. However, they require prior training and some knowledge about the identity of the speech or music sources in the mixture.

In this paper, we deal with the problem of extracting a speech source of interest from a known direction. we present a framework which extends the use of beamforming techniques to underdetermined speech mixtures. We describe frequency-domain non-linear mixture of beamformers that can extract a desired speech source from a known direction when there are fewer microphones than sources, and do not require knowledge of the number of speakers. These beamformers utilize Gaussian mixture models (GMMs) to model the observation data in each frequency bin. In contrast to other speech enhancement and separation methods which use GMMs such as [6], [23], [24], our approach do not couple the Gaussian states across frequency, and the covariance matrices of each Gaussian state represent a spatial covariance matrix. The model learning is performed using the observed mixture signals only, and no prior training is required. The signal estimator comprises of a set of MMSE or MVDR beamformers. In order to estimate the signal, all beamformers are concurrently applied to the observed signal, and the weighted sum of the beamformers' outputs is used as the signal estimator, where the weights are the posterior probabilities of the GMM states. These weights are specific to each time-frequency point. This approach results in a soft decision filter for the observed signal. The resulting non-linear beamformer combines the benefits of non-linear time-varying separation in time-frequency masking with the benefits of spatial filtering in the linear beamformers.

The remainder of the paper is structured as follows. Section II presents the signal mixing model. Section III reviews the linear MMSE, MVDR, and FMV beamformers. Then, in Section IV, the proposed GMM-based non-linear beamformers are described. The experimental conditions and simulation results are presented in Section V, followed by the conclusions in Section VI.

II. Mixing Model

Consider the convolved mixing model in (1). The time-domain observed signals $x_j(t)$ may be mapped to the time-frequency domain using the STFT. Denoting the STFT coefficients of $x_j(t)$ and $s_i(t)$ as $x_j(n,f)$ and $s_i(n,f)$, in the time frame n and frequency bin f, and approximating the mixing filters by complex mixing scalars $a_{ji}(f)$, we get:

$$x_j(n,f) = \sum_{i=1}^{M} a_{ji}(f) s_i(n,f) \tag{4}$$

Assuming we are only interested in extracting source i', $i' \in \{1,2,...,M\}$, the mixing model in (4)

can be reformulated as:

$$\begin{aligned} x_j(n,f) &= a_{ji'}(f)s_{i'}(n,f) + \sum_{\substack{i=1 \\ i\neq i'}}^{M} a_{ji}(f)s_i(n,f) \\ &= a_{ji'}(f)s_{i'}(n,f) + v_j(n,f) \end{aligned} \tag{5}$$

where v_j represents the contribution of the interferers to the mixture signal x_j. In vector form, the mixing model can be written as:

$$\mathbf{x}(n,f) = \mathbf{a}(f)s(n,f) + \mathbf{v}(n,f) \tag{6}$$

where $\mathbf{x}(n,f) = [x_1(n,f), ..., x_N(n,f)]^T$ is the observed multichannel mixture signal, $\mathbf{a}(f)$ is the $N \times 1$ array response vector in the direction of the desired source signal s (also called the propagation vector or steering vector), and $\mathbf{v}(n,f) = [v_1(n,f), ..., v_N(n,f)]^T$ is the $N \times 1$ vector of the interferers' contribution to the mixture signal. We assume that the direction of the desired source signal is known. In this model, no assumptions are made about the interferers. The interferers are not restricted to point sources in low reverberation conditions, but can also be of any nature such as spatial extended sources, diffuse sources, or a combination of them. The array response vector $\mathbf{a}(f)$ is the representation of the delays and the attenuation in the frequency domain, and depends on the array geometry and the direction of the desired source signal. If d were to represent the microphone spacing, c the sound velocity, ϕ the DOA relative to broadside, and assuming far-field conditions, we have for a uniform linear array:

$$\mathbf{a}(f) = [e^{-\iota 2\pi f \Delta_1}, ..., e^{-\iota 2\pi f \Delta_N}]^T \tag{7}$$

where $\iota = \sqrt{-1}$, and $\Delta_j = (j-1)(d/c)\sin\phi$. Note that $\mathbf{x}$, $\mathbf{a}$, s, and $\mathbf{v}$ are complex valued, and depend on frequency f, but for readability and simplicity, we will omit this variable in the rest of paper. From now on, we implicitly work in a given frequency band.

III. Optimum Beamformers

A. Linear MMSE Beamformer

We first consider the optimum estimator whose output is the MMSE estimate of the desired signal s in the presence of Gaussian interference, assuming a known desired signal direction. We assume that the desired source signal is a sample function from a zero-mean complex-valued Gaussian random process, $s \sim \mathbb{N}(0, \sigma_s^2)$. We also assume a zero-mean complex-valued Gaussian interference, $\mathbf{v} \sim \mathbb{N}(0, R_v)$. Additionally, it is assumed that the signal and interference snapshots are uncorrelated. Hence, $\mathbf{x} \sim$

$\mathbb{N}(0, R_v + \sigma_s^2 \mathbf{a}\mathbf{a}^H)$, and $\mathbf{x}|s \sim \mathbb{N}(\mathbf{a}s, R_v)$, where $(.)^H$ denotes the Hermitian transpose operator. The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\hat{s}_{\text{MMSE}} = \mathrm{E}\left[s|\mathbf{x}\right] = \int s \; p(s|\mathbf{x}) \; ds \tag{8}$$

This mean is referred to as the conditional mean. It can be shown that the conditional mean can be expressed as [25]:

$$\hat{s}_{\text{MMSE}} = \underbrace{\frac{\mathbf{a}^H R_v^{-1} \mathbf{x}}{\mathbf{a}^H R_v^{-1} \mathbf{a}}}_{\text{MVDR}} \underbrace{\frac{\sigma_s^2}{\sigma_s^2 + \left(\mathbf{a}^H R_v^{-1} \mathbf{a}\right)^{-1}}}_{\text{Wiener post filter}} \tag{9}$$

The first term is an MVDR spatial filter, which suppresses the interfering signals and noise without distorting the signal propagating along the desired source direction. The second term is a single-channel Wiener post-filter. We see that the MMSE estimator is just a shrinkage of the MVDR beamformer. Unfortunately, the MMSE beamformer depends explicitly on σ_s^2 which is typically unknown. However, we can obtain a beamformer that does not depend on σ_s^2 by imposing a distortionless response in the specified direction. The result is the MVDR beamformer, and the estimate of the desired source signal can be written as:

$$\hat{s}_{\text{MVDR}} = \frac{\mathbf{a}^H R_v^{-1}}{\mathbf{a}^H R_v^{-1} \mathbf{a}} \mathbf{x} \tag{10}$$

In general, the conditional mean estimator is not linear. The MMSE estimator is linear if either the estimator is constrained to be linear, or all the signals are Gaussian. However, speech sources are generally non-stationary and non-Gaussian. This suggests extending the optimum beamformers to exploit the non-stationarity and non-Gaussianity of speech signals.

B. Frequency-Domain MVDR (FMV) Beamformer

Speech is a non-stationary process, but over short durations speech signals can be considered stationary. In the FMV algorithm [10], it is assumed that source activity patterns are constant over small time intervals of speech signals in each frequency band, but could vary over longer time spans. In the FMV algorithm, time-frequency representations of the mixture signals are stored in a buffer, and a correlation matrix is calculated for each frequency bin using the 32 most recent mixture signal STFT values. MVDR weights are then calculated using the correlation matrix. Therefore, in the FMV algorithm, new beamformer weights are calculated for every small time interval, in order to reduce the contribution of interfering

sources active during that time interval to the extracted signal while maintaining a distortionless response in the desired source DOA. Only statistics gathered over a very short period of time are used in the calculation of weights.

The quick adaptation of the beamformer weights can substantially reduce a large number of non-stationary interferences while utilizing few microphones [10]. But the computational load is high due to recurrent matrix inversions in each frequency band and the need to have a very small step size in the STFT. In practice, however, source activity patterns can change abruptly between samples, and the FMV will perform spatial filtering based on the average power of the interfering sources active in the time interval during which the beamformer weights are calculated. On the other hand, the spatial distribution of the sources does not change very quickly, and we can gather statistics for the desired signal estimator over a longer time span. Thus the FMV beamformer is forced to compromise between long intervals (good statistics) and short intervals (rapid response).

IV. Proposed Method: Mixture of Beamformers

In the time-frequency domain, speech signals typically have a super-Gaussian (sparse) distribution, due to a combination of the non-stationarity and harmonic content of speech. Therefore, even if sources might overlap at some time-frequency points, not all speech sources in a mixture are active at the same time-frequency points. It is therefore advantageous to exploit the sparsity property of speech signals in the time-frequency domain in order to perform separation in underdetermined environments. In this paper, we use GMMs to model the speech non-Gaussianity and the spatial distribution of the sources.

In this section, we present three non-linear beamformers that can perform underdetermined speech separation. The first two non-linear beamformers are based on modeling the desired source signal s and the interference $\mathbf{v}$ separately. The desired source signal is modeled using a 1-dimensional GMM, and the observed interference is modeled using an N-dimensional GMM, where N is the number of mixture channels. The third non-linear beamformer is based on modeling the observed mixture signal $\mathbf{x}$ (the desired source and interference together) using an N-dimensional GMM.

We describe the density of the interference signal $\mathbf{v}$ in each frequency bin as a mixture of k_v zero-mean, complex-valued, N-dimensional Gaussians with indices $q_v = 1, ..., k_v$, covariances R_{v,q_v} and mixing proportions c_{v,q_v}:

$$p(\mathbf{v}|\theta_v) = \sum_{q_v=1}^{k_v} c_{v,q_v} \frac{1}{\pi^N |R_{v,q_v}|} \exp\left(-\mathbf{v}^H R_{v,q_v}^{-1} \mathbf{v}\right) \tag{11}$$

where $\theta_v = \{c_{v,q_v}, R_{v,q_v} : 1 \leq q_v \leq k_v\}$, and the mixing proportions $c_{v,q_v} = p(q_v)$ (prior probabilities of the Gaussian states) are constrained to sum to one. In addition, we shall describe the density of the desired source signal s in each frequency bin as a mixture of k_s zero-mean complex-valued 1-dimensional Gaussians with indices $q_s = 1, ..., k_s$, variances σ^2_{s,q_s} and mixing proportions c_{s,q_s}:

$$p(s|\theta_s) = \sum_{q_s=1}^{k_s} c_{s,q_s} \frac{1}{\pi\sigma^2_{s,q_s}} \text{exp}\left(\frac{-|s|^2}{\sigma^2_{s,q_s}}\right) \tag{12}$$

where $\theta_s = \{c_{s,q_s}, \sigma^2_{s,q_s} : 1 \leq q_s \leq k_s\}$, and the mixing proportions $c_{s,q_s} = p(q_s)$ (prior probabilities of the Gaussian states) are constrained to sum to one. The number of components k_s and k_v control the flexibility of the model. In our model, the Gaussian states are not coupled across frequency, and the parameters $\{\theta_s, \theta_v\}$ are frequency dependent.

The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\begin{aligned}
\hat{s}_{\text{MMSE}} &= \text{E}\left[s|\mathbf{x}\right] = \int p(s|\mathbf{x})\ s\ ds \\
&= \int \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(s, q_s, q_v|\mathbf{x})\ s\ ds \\
&= \int \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(q_s, q_v|\mathbf{x})\ p(s|\mathbf{x}, q_s, q_v)\ s\ ds \\
&= \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(q_s, q_v|\mathbf{x}) \int p(s|\mathbf{x}, q_s, q_v)\ s\ ds \\
&= \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \text{E}\left[s|\mathbf{x}, q_s, q_v\right]
\end{aligned} \tag{13}$$

where

$$\begin{aligned}
\tau_{q_s,q_v} &= p(q_s, q_v|\mathbf{x}) \\
&= \frac{p(\mathbf{x}|q_s, q_v)\ p(q_s)\ p(q_v)}{\sum_{q_s'=1}^{k_s} \sum_{q_v'=1}^{k_v} p(\mathbf{x}|q_s', q_v')\ p(q_s')\ p(q_v')}
\end{aligned} \tag{14}$$

is the a posteriori probability that the components q_s and q_v are active in their respective GMMs when observing $\mathbf{x}$, with $\sum_{q_s} \sum_{q_v} \tau_{q_s,q_v} = 1$. The posteriori probability is specific to each time frequency point, and has a non-linear dependency on the observed data.

We can see that the conditional mean $\text{E}\left[s|\mathbf{x}, q_s, q_v\right]$ is the linear MMSE beamformer estimator in (9), with $R_v = R_{v,q_v}$ and $\sigma^2_s = \sigma^2_{s,q_s}$. The desired signal estimator in (13) is a non-linear weighted sum of linear MMSE beamformers over all the GMM components, and the weighting coefficients are the

a posteriori probabilities of the GMM components τ_{q_s,q_v} (specific to each time-frequency point). This mixture of MMSE beamformers will be denoted by $\mathbf{w}_1$ and is given by [26]:

$$\mathbf{w}_1 = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \frac{\sigma_{s,q_s}^2}{\sigma_{s,q_s}^2 + \left(\mathbf{a}^H R_{v,q_v}^{-1} \mathbf{a}\right)^{-1}} \frac{\mathbf{a}^H R_{v,q_v}^{-1}}{\mathbf{a}^H R_{v,q_v}^{-1} \mathbf{a}} \tag{15}$$

In comparison to independent factor analysis [27], where sources were also modeled with GMMs, the mixture of MMSE beamformers models all the interfering sources using one N-dimensional mixture of Gaussians in the observation (microphones) domain. Consequently, the number of interferers in the mixture is not required to be known or have a unique mixing structure. This also avoids the exponential growth of the number of Gaussian components in the observation density with the number of sources.

If a distortionless response in the direction of the desired source is required, a distortionless response mixture of MVDR beamformers can be used [26]:

$$\mathbf{w}_2 = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \frac{\mathbf{a}^H R_{v,q_v}^{-1}}{\mathbf{a}^H R_{v,q_v}^{-1} \mathbf{a}} \tag{16}$$

This mixture of MVDR beamformers is a non-linear weighted sum of linear distortionless MVDR beamformers, where the weights sum to unity. As a result, it is constrained to a distortionless response in the look-direction. By distortionless we mean it has a unity gain in the look-direction at all time-frequency points.

In Appendix A, we develop an EM algorithm to learn the model density parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma_{s,q_s}^2, c_{v,q_v}, R_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$. The model learning is applied separately in each frequency bin.

We briefly summarize the main steps in the separation procedure using $\mathbf{w}_1$ or $\mathbf{w}_2$ in Algorithm 1. Note that the model learning step is applied separately in each frequency bin, and that the Gaussian states' posterior probabilities are specific to each time-frequency point (no coupling across all frequencies).

In a previous paper [28], a non-linear beamformer was developed assuming a distortionless response in the direction of the desired source. A mixture of k_x zero-mean, complex-valued, N-dimensional Gaussians with indices $q_x = 1, ..., k_x$, covariances R_{x,q_x} and mixing proportions c_{x,q_x} was used to model the observed mixture $\mathbf{x}$ (the desired source and interference together) in each frequency bin:

$$p(\mathbf{x}|\theta_x) = \sum_{q_x=1}^{k_x} c_{x,q_x} \frac{1}{\pi^N \left|R_{v,q_v}\right|} \exp\left(-\mathbf{x}^H R_{x,q_x}^{-1} \mathbf{x}\right) \tag{21}$$

where $\theta_x = \{c_{x,q_x}, R_{x,q_x} : 1 \leq q_x \leq k_x\}$, and the mixing proportions $c_{x,q_x} = p(q_x)$ (prior probabilities of the Gaussian states) are constrained to sum to one. This leads to a simple learning algorithm, and the

Algorithm 1 Separation procedure using $\mathbf{w}_1$ or $\mathbf{w}_2$

1: Compute the STFT of the mixture $\mathbf{x}$.

2: Apply the EM algorithm (see Appendix A) separately in each frequency bin to compute $\{\tau_{q_s,q_v}(n,f)$, $\sigma^2_{s,q_s}(f), R_{v,q_v}(f) : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$.

3: For each time-frequency point $(n,\ f)$, the output of the non-linear beamformer is given by:

$$\hat{s}(n,f) = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}(n,f)\, \mathbf{w}_{q_s,q_v}(f)\, \mathbf{x}(n,f) \tag{17}$$

where $\mathbf{w}_{q_s,q_v}(f)$ can be either a linear MMSE or a linear MVDR beamformer:

$$\mathbf{w}^{\text{MVDR}}_{q_s,q_v}(f) = \frac{\mathbf{a}(f)^H R^{-1}_{v,q_v}(f)}{\mathbf{a}(f)^H R^{-1}_{v,q_v}(f)\mathbf{a}(f)} \tag{18}$$

$$\mathbf{w}^{\text{MMSE}}_{q_s,q_v}(f) = H^{\text{Wiener}}_{q_s,q_v}(f)\, \mathbf{w}^{\text{MVDR}}_{q_s,q_v}(f) \tag{19}$$

where the scalar, single channel Wiener post filter is given by:

$$H^{\text{Wiener}}_{q_s,q_v}(f) = \frac{\sigma^2_{s,q_s}(f)}{\sigma^2_{s,q_s}(f) + \left(\mathbf{a}(f)^H R^{-1}_{v,q_v}(f)\mathbf{a}(f)\right)^{-1}} \tag{20}$$

4: The corresponding time-domain signal $\hat{s}$ is derived by an STFT inversion.

learning of model parameters is independent on the desired source direction. The desired signal can be estimated using this mixture of MVDR beamformers [28]:

$$\mathbf{w}_3 = \sum_{q_x=1}^{k_x} \tau_{q_x} \frac{\mathbf{a}^H R^{-1}_{x,q}}{\mathbf{a}^H R^{-1}_{x,q}\mathbf{a}} \tag{22}$$

where $\tau_{q_x} = p(q_x|\mathbf{x})$ is the relative contribution for each linear MVDR beamformer, and is calculated as the posterior probability (specific to each time-frequency point) of its corresponding Gaussian component. The resulting beamformer has a unity gain in the look-direction at all time-frequency points.

In Appendix B, we develop an EM algorithm to learn the observation model density parameters $\theta_x = \{c_{x,q_x}, R_{x,q_x} : 1 \leq q_x \leq k_x)$. The model learning is applied separately in each frequency bin.

The main steps in the separation procedure using $\mathbf{w}_3$ are summarized in Algorithm 2.

V. Experimental Evaluation

A. Setup

In order to illustrate the performance of the non-linear beamformers, multichannel recordings of several speech sources were simulated using impulse responses determined by the room image method [29]. The

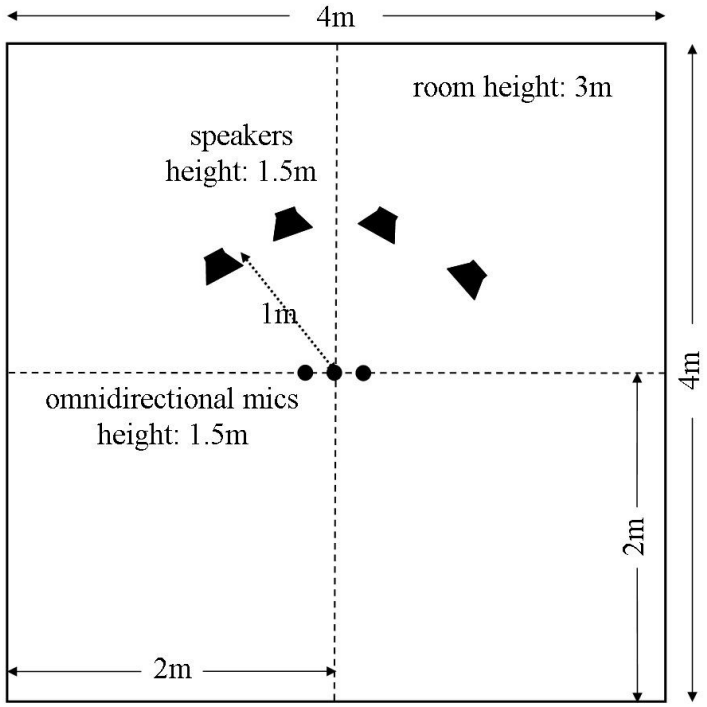

Fig. 1. Layout of room used in simulations.

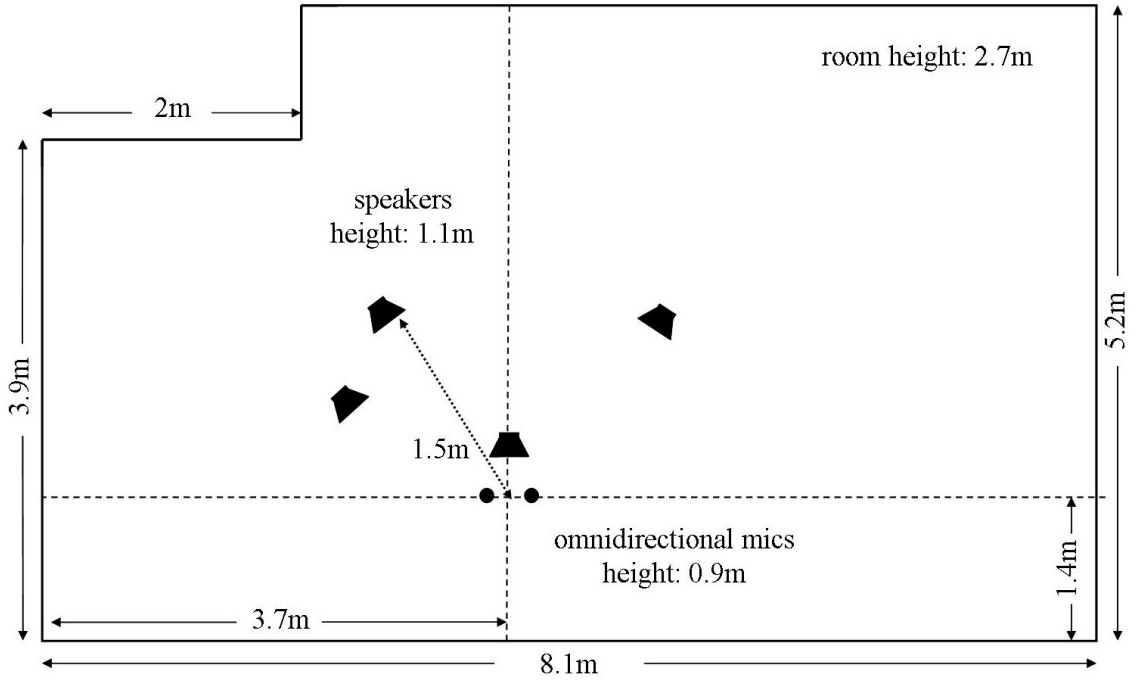

Fig. 2. Layout of room used in recordings

Algorithm 2 Separation procedure using $\mathbf{w}_3$

1: Compute the STFT of the mixture $\mathbf{x}$.

2: Apply the EM algorithm (see Appendix B) separately in each frequency bin to compute $\{\tau_{q_x}(n,f), R_{x,q_x}(f) : 1 \leq q_x \leq k_x\}$.

3: For each time-frequency point $(n,\ f)$, the output of the non-linear beamformer is given by:

$$\hat{s}(n,f) = \sum_{q_x=1}^{k_x} \tau_{q_x}(n,f)\, \mathbf{w}_{q_x}(f)\, \mathbf{x}(n,f) \tag{23}$$

where:

$$\mathbf{w}_{q_x}(f) = \frac{\mathbf{a}(f)^H R_{x,q_x}^{-1}(f)}{\mathbf{a}(f)^H R_{x,q_x}^{-1}(f)\, \mathbf{a}(f)} \tag{24}$$

4: The corresponding time-domain signal $\hat{s}$ is derived by an STFT inversion.

positions of the microphones and the sources were as illustrated in Fig. 1. Two microphone arrays were used. The first has three microphones with a spacing $d = 2.5$ cm, and the second has two microphones with a spacing $d = 5$ cm. In section V-H, live recordings in the room illustrated in Fig. 2 were used.

We used speech files taken from the TIMIT speech corpus [30] to create five mixtures of male sources, and five mixtures of female sources. The speech signals were of a duration equal to 10 s, and were sampled at 16 kHz. The number of the sources in each mixture was four. The sources were placed in a semi-circle of radius 1 m around the microphone arrays at angles $\phi = \{-45,\ -15,\ 10,\ 50\}^\circ$.

B. Evaluation Measures

To measure the quality of the signal estimate $\hat{s}$ with respect to the original signal s, we used the source to distortion ratio (SDR), source to interference ratio (SIR) and the sources to artifacts ratio (SAR) calculated as defined in [31]. The computation of the evaluation measures involves two steps. First, the estimated signal $\hat{s}$ is decomposed as

$$\hat{s} = s_{\text{target}} + e_{\text{interf}} + e_{\text{artif}} \tag{25}$$

where s_{target} is a version of the desired source s modified by an allowed distortion, and where e_{interf} and e_{artif} are respectively the interferences and artifacts error terms. In a second step, we compute energy

TABLE I

ALGORITHM PARAMETERS

		$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$
STFT frame		1024	1024	1024
STFT step		256	256	256
GMM components	(2 mics)	$k_s = 2$, $k_v = 15$	$k_s = 2$, $k_v = 15$	$k_x = 15$
	(3 mics)	$k_s = 2$, $k_v = 5$	$k_s = 2$, $k_v = 5$	$k_x = 5$
EM Iterations	(2 mics)	100	100	50
	(3 mics)	100	100	20

ratios to evaluate the relative amount of each of these terms as follows:

$$\text{SDR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artif}}\|^2} \tag{26}$$

$$\text{SIR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2} \tag{27}$$

$$\text{SAR} = 10\log_{10}\frac{\|s_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2} \tag{28}$$

In our results, the SDR, SIR and SAR values were averaged over all the sources and mixtures.

C. Algorithm Parameters

Unless mentioned otherwise, we use the values listed in Table I for the STFT frame size, STFT step size, number of GMM components, and number of iterations.

D. Effect Of Design Parameters

We first investigate the effect of various parameters on the performance of the non-linear beamformers. We study the effect of the number of Gaussian components in the GMM model, the required number of EM iterations, and the effect of the learning block size.

1) Effect of the Number of Gaussian Components: In this experiment, four sources were operating in an anechoic environment (RT = 0), and the microphone array used has two microphones with a 5 cm microphone spacing. Fig. 3 shows the average performance at the output of the non-linear beamformer $\mathbf{w}_3$ defined in (22) as a function of the number of Gaussian components k_x in the GMM model. The case of $k_x = 1$ is equivalent to a time-invariant MVDR beamformer. The SIR increases with k_x, but the improvement is insignificant at $k_x > 10$. Although there is a unity-gain response in the direction

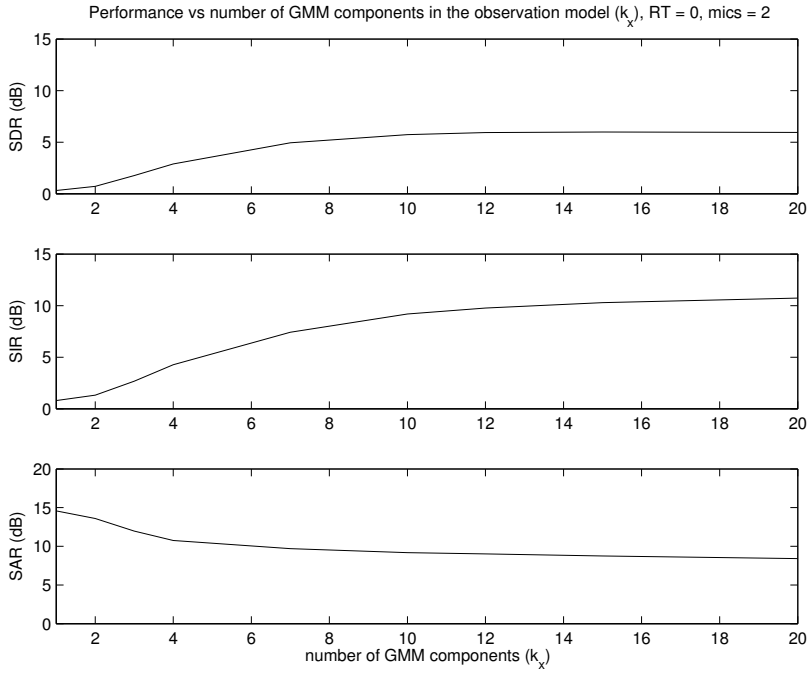

Fig. 3. Average performance of the non-linear beamformer $\mathbf{w}_3$ in equation (22) as a function of the number of Gaussian components k_x in the GMM model.

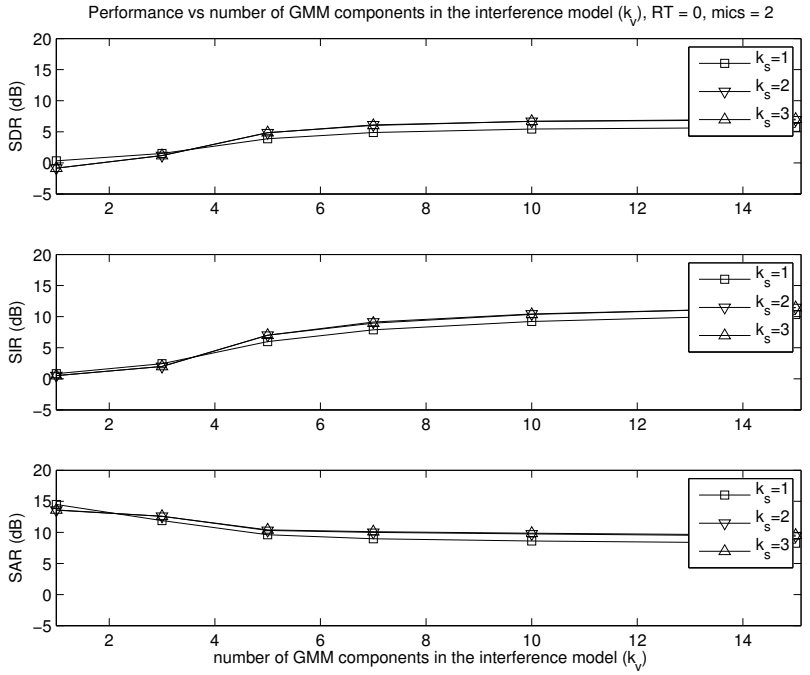

Fig. 4. Average performance of the non-linear beamformer $\mathbf{w}_2$ in equation (16) as a function of the number of Gaussian components k_v and k_s in the GMM model.

of the desired source signal, the SAR decreases with k_x. The decrease in the SAR can be attributed to the non-linear attenuation of the interfering sources. These artifacts therefore introduce distortion only into the residual interfering signals. We stress that the mixture of MVDR beamformers is by definition distortionless in the look-direction.

Fig. 4 shows the average performance at the output of the non-linear beamformer $\mathbf{w}_2$ defined in (16) as a function of the number of Gaussian components in the interference model k_v and the number of Gaussian components in the source model k_s. We can see that there is little to be gained in increasing the number of source Gaussian components k_s to more than two. The SIR increases with k_v, but the improvement again is insignificant for $k_v > 10$. The non-linear beamformer can attain a SIR of 10 dB in the two microphones case.

Fig. 5 shows the average performance at the output of the mixture of MMSE beamformers $\mathbf{w}_1$ defined in (15) as a function of the number of Gaussian components in the interference model k_v and the number of Gaussian components in the source model k_s. The non-linear beamformer can attain an SIR of 13 dB. However, the SAR was reduced in comparison to Fig. 4 because the distortionless constraint is no longer held.

2) Effect of the Number of Iterations: Fig. 6 shows the average performance at the output of the non-linear beamformers in the anechoic case as a function of the number of EM iterations. The microphone array used has two microphones with a 5 cm microphone spacing. The non-linear beamformer $\mathbf{w}_3$ defined in (22) require less than 20 iterations to converge, whereas the other two non-linear beamformers require more iterations to converge (about 100 iterations).

3) Effect of the Learning Block Size: The EM algorithm used in our experiments is a batch learning algorithm. We studied the effect of varying the size of learning data on the performance of the non-linear beamformers. Fig. 7 shows the average performance at the output of the non-linear beamformer $\mathbf{w}_3$ defined in (22) in the anechoic case as a function of the EM learning block length. The performance is fairly consistent even when using shorter learning blocks. Note that the FMV algorithm can be considered as a special case of the non-linear beamformer $\mathbf{w}_3$, with $k_x = 1$ and very short learning blocks (≈ 100 ms).

E. Directivity Patterns

Fig. 8 shows four examples of directivity patterns for the non-linear beamformer $\mathbf{w}_3$ of equation (22) in the anechoic case. The directivity patterns are defined as the magnitude of the response of the beamformer

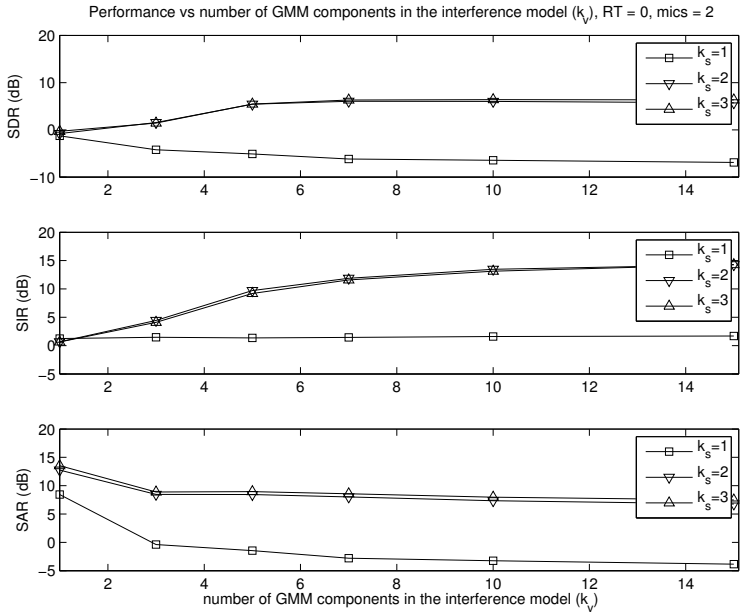

Fig. 5. Average performance of the non-linear beamformer $\mathbf{w}_1$ in equation (15) as a function of the number of Gaussian components k_v and k_s in the GMM model.

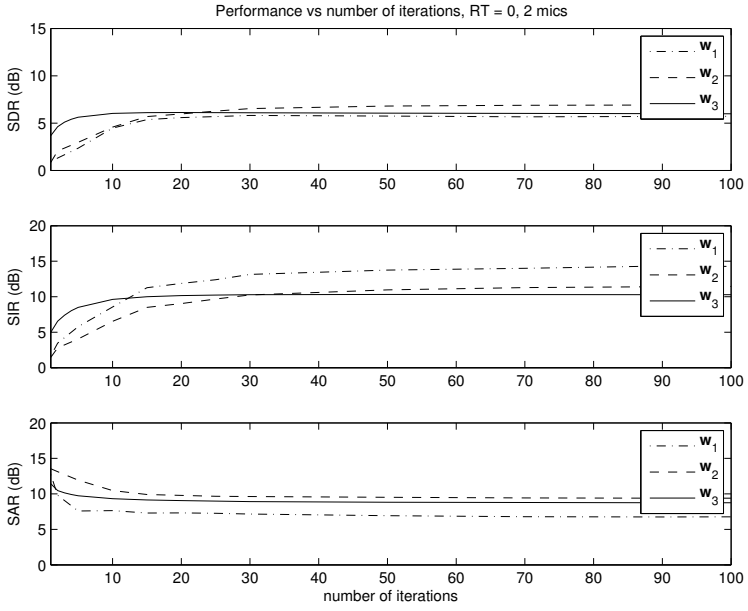

Fig. 6. Separation using two microphones: average performance as a function of the number of EM iterations.

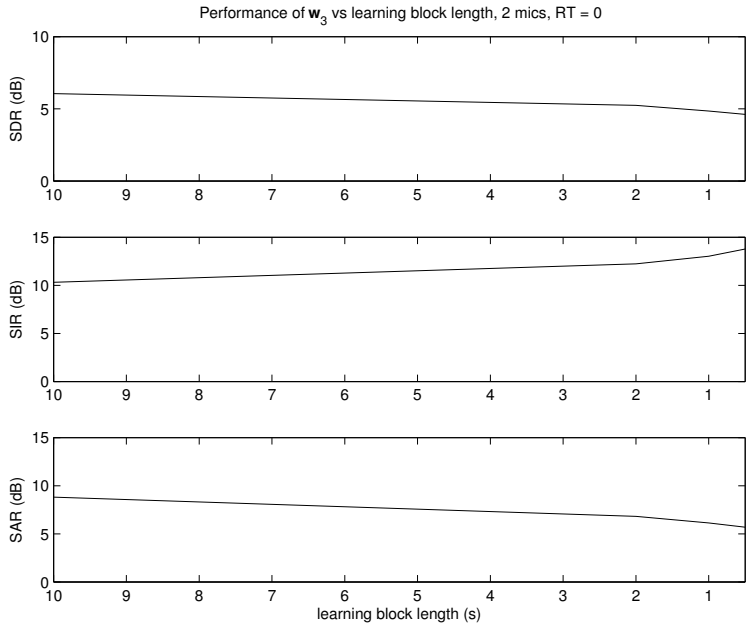

Fig. 7. Average performance of $\mathbf{w}_3$ vs learning block length in seconds

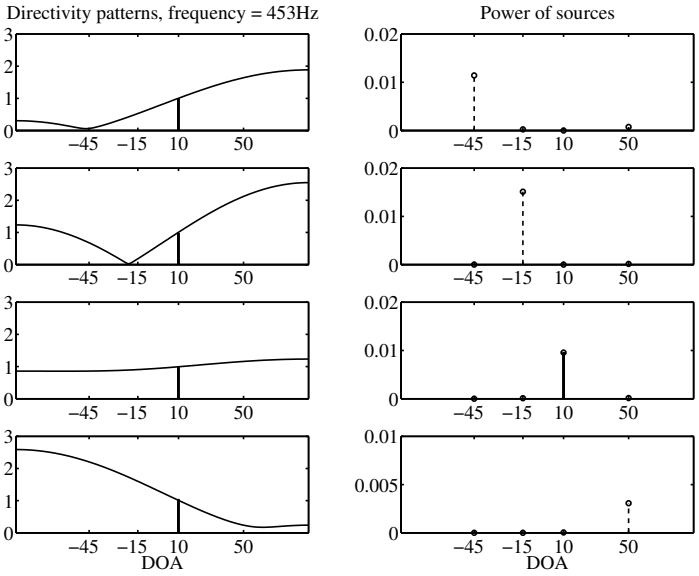

Fig. 8. Examples of directivity patterns at 453 Hz, when the desired source is at angle 10°, and the interfering sources are at $\{-45,\ -15,\ 50\}^\circ$. Left column: directivity patterns. Right column: power of sources.

at frequency f for a far-field signal coming from direction Φ:

$$\mathrm{D}(f,\Phi) = \left| \sum_{j=1}^{N} w_j(f).e^{\iota 2\pi f(j-1)dc^{-1}\sin\Phi} \right| \tag{29}$$

In this experiment, the desired source was at an angle of 10°, and the interfering sources at $\{-45,\ -15,\ 50\}^\circ$. The microphone array used has two microphones with a 5 cm microphone spacing. The four examples are at four different time frames at the frequency of 453 Hz. In the first example (first row), the desired source and the interferer at angle -45° were active. In the second example (second row), the interferer at angle -15° was active. In the third example (third row), the desired source was active, and in the fourth example (fourth row), the interferer at angle 50° was active. The non-linear beamformer effectively nullifies the active interferer while having a distortionless response in the direction of the desired source.

F. Effect Of DOA Offset

In a typical application, the DOA of the desired source is scanned across a region of interest in space. The desired signal can arrive from a different direction than that assumed. We tested the effect of the mismatch between the assumed DOA of the desired source and the true one. Fig. 9 shows the average performance at the output of the non-linear beamformers in the anechoic case as a function of the DOA offset. The non-linear beamformers appear to be robust to small DOA offsets.

G. Effect Of Reverberation

Fig. 10 shows the average performance as a function of the room reverberation time when four sources are operating, and the microphone array used has two microphones with a 5 cm microphone spacing. $k_x = 15$ was used in the beamformer defined in (22), and $k_s = 2, k_v = 15$ was used in the two other beamformers. We compared the performance of the three non-linear beamformers with the performance of the MENUET and FMV algorithms. A STFT of frame size 1024 samples is used. In the FMV algorithm, a small step size of 16 samples is required, while a step size of 256 samples is sufficient in the non-linear beamformers. The MENUET algorithm and the mixture of MMSE beamformers ($\mathbf{w}_1$) gives a high SIR, but suffers from a very low SAR at higher reverberation times. The non-linear beamformers $\mathbf{w}_2$ and $\mathbf{w}_3$ of equations (16) and (22) respectively have significantly lower artifacts at higher reverberation times.

Fig. 11 shows the average performance as a function of the room reverberation time when four sources are operating, and the microphone array used has three microphones with a 2.5 cm microphone spacing. We compared the performance of the three non-linear beamformers with the performance of the MENUET

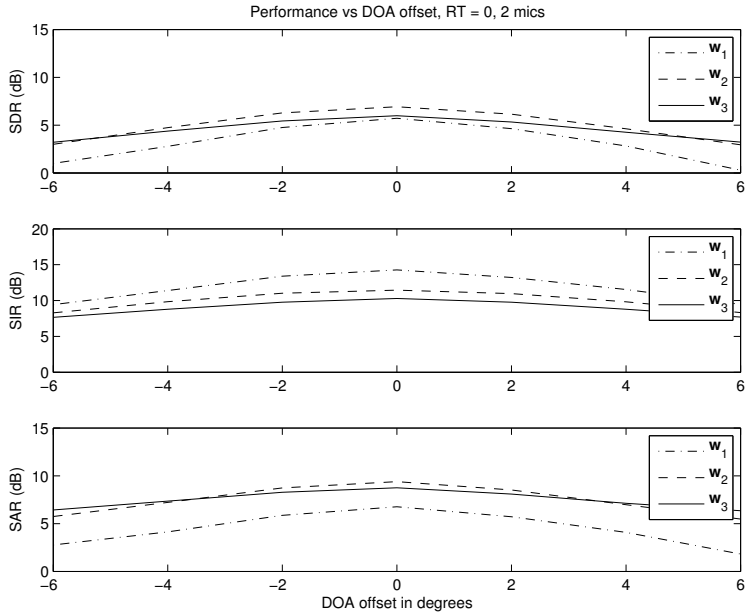

Fig. 9. Separation using two microphones: average performance as a function of DOA offset.

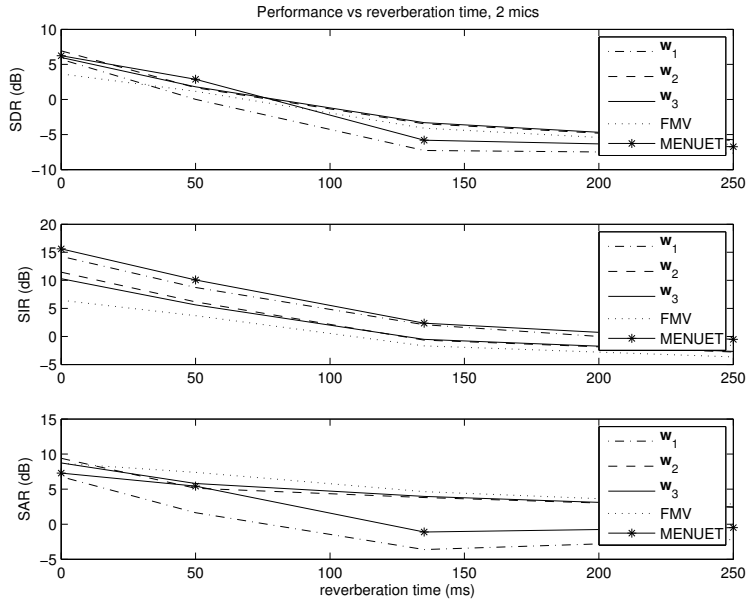

Fig. 10. Separation using two microphones: average performance as a function of reverberation time.

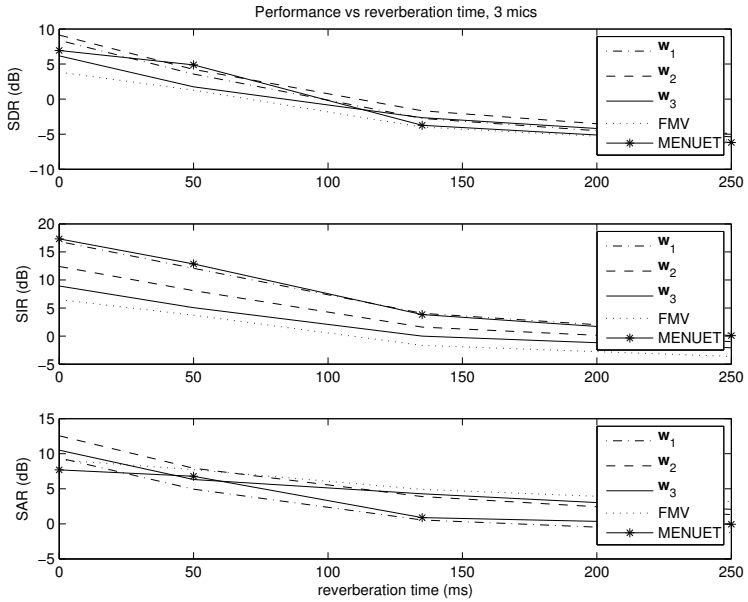

Fig. 11. Separation using three microphones: average performance as a function of reverberation time.

and FMV algorithms. $k_x = 5$ was used in the beamformer of equation (22), and $k_s = 2, k_v = 5$ was used in the two other beamformers. The performance of non-linear beamformers improved with the addition of the third microphone.

H. Live Recordings

In order to illustrate the performance of the non-linear beamformers in real life recordings, multichannel recordings of several speech sources were recorded in a room with a reverberation time of 810 ms. The dimensions of the room and the positions of the microphones and the sources are illustrated in Figure 2. The microphone array has two microphones with spacing $d = 7$ cm. We use the same speech files used in the simulations (five mixtures of male sources, and five mixtures of female sources). The number of the sources in each mixture was four. The desired source was placed 30 cm away from the microphone array, while the interferers were placed in a semi-circle of radius 1.5 m around the microphone arrays at angles $\phi = \{-60, -30, 50\}^\circ$. We compared the three non-linear beamformers with the FMV and MENUET algorithms. The SIR and SAR values were averaged over all the mixtures. Table II shows the results.

Due to the high reverberation times, all of the methods suffer from low SIR values, but they all afford SIR improvements over the input mixture (the mixture SIR is –4.7 dB). The non-linear beamformer $\mathbf{w}_2$

TABLE II

AVERAGE PERFORMANCE USING REAL LIFE RECORDINGS IN A ROOM WITH 810 MS REVERBERATION TIME

	SDR	SIR	SAR
$\mathbf{w}_1$	0.3	3.8	4.5
$\mathbf{w}_2$	-1.7	-0.5	7.9
$\mathbf{w}_3$	-1.8	-0.3	7.0
FMV	-2.1	-0.3	5.7
MENUET	-0.4	3.3	3.6

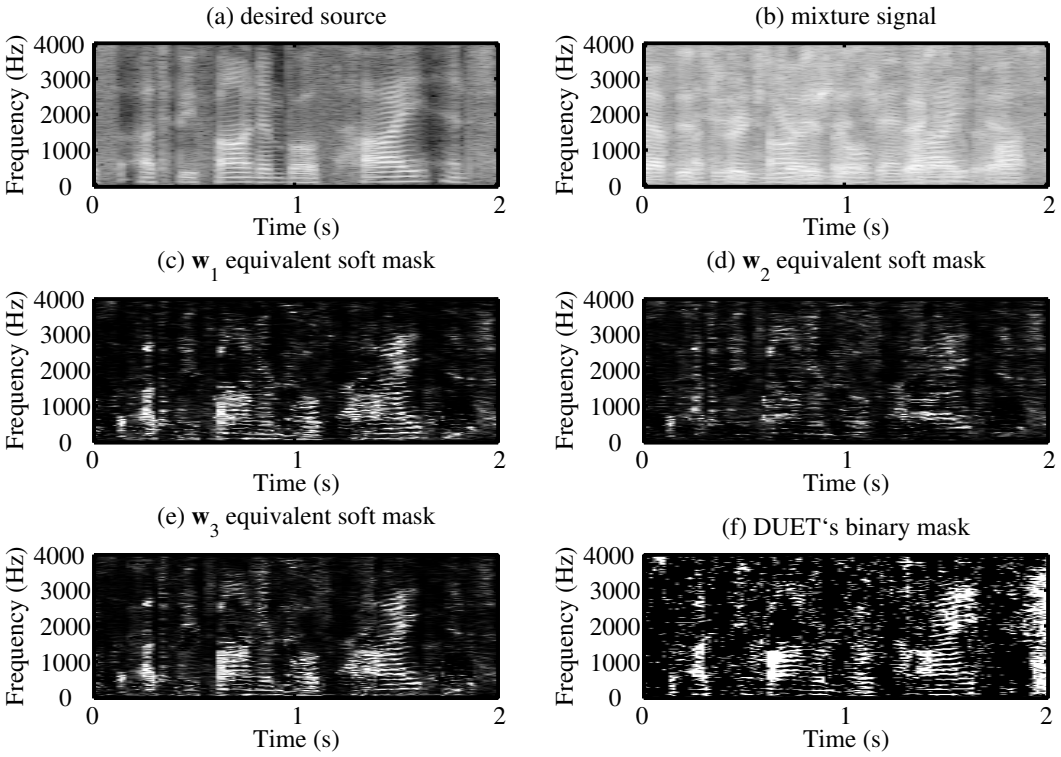

Fig. 12. (a) Spectrogram of a desired signal. (b) Spectrogram of a mixture of 4 sources. (c-f) t-f masks.

shows the highest SAR. The non-linear beamformer $\mathbf{w}_1$ has the highest SIR and SDR performance, and also achieves better SAR than MENUET which had the second best SIR.

TABLE III

COMPUTATIONAL TIME FOR THE PROPOSED METHODS

	time (s)	parameters
$\mathbf{w}_1$ (2 mics)	670	$k_s = 2$, $k_v = 15$, 100 iterations
$\mathbf{w}_2$ (2 mics)	607	$k_s = 2$, $k_v = 15$, 100 iterations
$\mathbf{w}_3$ (2 mics)	172	$k_x = 15$, 50 iterations
$\mathbf{w}_1$ (3 mics)	248	$k_s = 2$, $k_v = 5$, 100 iterations
$\mathbf{w}_2$ (3 mics)	230	$k_s = 2$, $k_v = 5$, 100 iterations
$\mathbf{w}_3$ (3 mics)	41	$k_x = 5$, 20 iterations

I. Time-Frequency Masks

To understand how the various beamformers are achieving their signal enhancement, we can look at equivalent time-frequency masks for each algorithm. Fig. 12 compares the equivalent mask of the three non-linear beamformers with the time-frequency mask of DUET on an example mixture. The equivalent mask was computed at each time-frequency point as the ratio of the energy of the desired signal estimate to the energy of the observed mixture. The non-linear beamformer's approach results in a soft decision mask for the observed signal.

J. Computational Time

In this subsection, we report the time it took for our Matlab implementation of the non-linear beamformers to run on 2.5 GHz CPU. The time reported is for the extraction of one 10 s speech source. We used the same design parameters used in subsection V-G. Table III shows the results.

We can see that the beamformer $\mathbf{w}_3$ took less time than the other two beamformers. Beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$ have comparable computational time because they use the same learning algorithm (Appendix A). The computational time for the three microphones case is lower than that of the two microphones case. This is because the performance of the beamformers using the three microphones array peak using fewer Gaussian components than the two microphone case.

VI. CONCLUSION

Frequency-domain non-linear mixture of beamformers were introduced and applied to the extraction of a desired speech source from a known direction in underdetermined speech mixtures. The system model assumes an anechoic desired source signal, but no assumptions are made about the interferers, which can

be of any nature such as point sources, spatial extended sources, diffuse sources, or a combination of them. The beamformers are derived assuming non-Gaussian interference signals modeled using a mixture of Gaussians distribution. This estimator introduces additional degrees of freedom to the beamformer by exploiting the super-Gaussianity (sparsity) of the interferers and dynamically steers nulls in order to reduce active interfering signals.

The non-linear beamformers require the location of the target speech source to be known or estimated in advance, but they have the following advantages:

- No need to know - or estimate - the number of interfering sources.
- Can be applied to underdetermined speech mixtures.
- The number of components in the GMM model controls the flexibility of the model. We did not incur overfitting in our experiments, therefore the number of GMM components can be used to trade-off complexity with performance. When using a larger number of microphones, the performance peaks with a small number of GMM components.
- Can be applied to microphone arrays with two or more microphones.
- Robust to small errors in the desired source DOA.

While one could impose models that are coupled across frequency to represent spectral patterns, we want to avoid that to keep the model as general as possible. This allows a close match to the actual properties of the observed signals and avoids the effect of microphone and channel variability which can cause a mismatch with the prior training conditions [23]. With our GMM model, we are aiming to impose as little structure on the source and interference models as possible. However, in future work, we would like to investigate the effect of source specific models.

The non-linear beamformers have been tested and evaluated on underdetermined speech mixtures. It was shown that the non-linear beamformer $\mathbf{w}_1$ defined in (15) gives better interference rejection at the expense of higher artifacts, especially at higher reverberation times. The non-linear beamformers $\mathbf{w}_2$ and $\mathbf{w}_3$ defined in (22) and (16) are distortionless beamformers (constant gain in the look-direction), and have significantly lower artifacts at higher reverberation times.

In terms of computational complexity, non-linear beamformer $\mathbf{w}_3$ employs a simpler learning algorithm and require fewer iterations than non-linear beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$. Furthermore, the model learning for non-linear beamformer $\mathbf{w}_3$ is independent of the location of the desired source, which makes this non-linear beamformer suitable in applications where scanning for the source direction is needed.

In our current implementation, the EM algorithm used is in a batch learning mode. In section V.D.3, we studied the effect of using short blocks of data. The batch mode with short blocks of data can be used

in applications where short delays are permissible, such as in human-computer interaction or surveillance. However, it is not appropriate for real-time applications. In these applications, online model learning is essential [32]. The online model learning should have a forgetting factor, and a mechanism for adding, deleting, and reassigning Gaussians to handle changes in the environment [33].

In the future, we would like to investigate the use of other linearly constrained minimum variance (LCMV) beamformers and Bayesian beamformers that are robust to DOA uncertainty [34] in the mixture of beamformers framework. We would also like to investigate the use of other filter banks instead of the STFT, such as auditory or constant-Q filter banks [35]. Through this, we aim to improve the performance of the beamformers at higher reverberation times.

APPENDIX
DERIVATION OF THE EM ALGORITHM

Using the EM algorithm, we can estimate the model density parameters from a set of observations $D = \{\mathbf{x}(n) : 1 \leq n \leq \eta\}$. The EM algorithm is used to find a maximum likelihood estimate of parameters in probabilistic models with latent variables (incomplete data problems). In our case, $\mathbf{x}$ is the observed (or incomplete) data, and the latent variables are the state sequence of the Gaussian mixtures that indicate which Gaussian components are responsible for $\mathbf{x}(n)$. In EM terminology, the complete data is composed of both the observed data and the latent variables. The EM algorithm is an iterative algorithm with two steps: (1) an expectation step (E-step), and (2) a maximization step (M-step). In the E-step, we calculate the conditional expectation of the complete data likelihood. The expectation is taken with respect to the conditional probability of the hidden states, given the observed data and the parameter values obtained in the previous iteration. In the M-step, the new estimates of the parameters are calculated to maximize the conditional expectation of the complete data likelihood.

A. Learning Interference And Desired Source Parameters

In this section, the parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma^2_{s,q_s}, c_{v,q_v}, R_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$ of the interference $\mathbf{v}$ and desired source s are estimated using the EM algorithm. These parameters are required for the non-linear beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$ of equations (15) and (16). Let us define a complete data set $D_c = \{\mathbf{x}, s, q_s, q_v\}$ composed of both the observed and the latent data. If we were to actually have such a complete data set, we could define its likelihood as:

$$l_c(\theta|D_c) = \ln \prod_{n=1}^{\eta} p(\mathbf{x}(n), s(n), q_s(n), q_v(n)|\theta) = \sum_{n=1}^{\eta} \ln p(\mathbf{x}(n), s(n), q_s(n), q_v(n)|\theta) \tag{30}$$

Given an initial value θ^0, the EM algorithm performs the following steps at each iteration l:

E-step:: In the E-step, we compute the expectation of the complete data likelihood:

$$Q(\theta, \theta^{l-1}) = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \int ds\, p\left(s, q_s, q_v | \mathbf{x}, \theta^{l-1}\right) \ln p(\mathbf{x}, s, q_s, q_v | \theta) \tag{31}$$

In the Gaussian mixture problem, this simply reduces to calculating $p\left(q_s, q_v | \mathbf{x}, \theta^{l-1}\right)$, the posterior probability of the latent variables, given the observed data and the parameters obtained in the previous iteration.

$$\begin{aligned} \tau_{q_s,q_v}^{(l)} &= p\left(q_s, q_v | \mathbf{x}, \theta^{(l-1)}\right) \\ &= \frac{p\left(q_s, q_v, \mathbf{x} | \theta^{(l-1)}\right)}{p\left(\mathbf{x} | \theta^{(l-1)}\right)} \\ &= \frac{p\left(q_s, q_v | \theta^{(l-1)}\right) p\left(\mathbf{x} | q_s, q_v, \theta^{(l-1)}\right)}{\sum_{q_s'=1}^{k_s} \sum_{q_v'=1}^{k_v} p\left(q_s', q_v' | \theta^{(l-1)}\right) p\left(\mathbf{x} | q_s', q_v', \theta^{(l-1)}\right)} \end{aligned} \tag{32}$$

where

$$\begin{aligned} p(\mathbf{x} | q_s, q_v) &= \int p(\mathbf{x}, s | q_s, q_v)\ ds \\ &= \int p(\mathbf{x} | s, q_v)\ p(s | q_s)\ ds \\ &= \int \mathbb{N}\left(\mathbf{x} - \mathbf{a}s, R_{v,q_v}\right)\ \mathbb{N}\left(s, \sigma_{s,q_s}^2\right)\ ds \\ &= \mathbb{N}\left(\mathbf{x}, R_{v,q_v} + \sigma_{s,q_s}^2 \mathbf{a}\mathbf{a}^H\right) \end{aligned} \tag{33}$$

Moreover, we evaluate the conditional mean and variance of the desired source given both the observed mixture and the hidden states, which are denoted by $\langle s | \mathbf{x}(n), q_s, q_v \rangle$ and $\langle ss^* | \mathbf{x}(n), q_s, q_v \rangle$ respectively. Given the hidden states and the mixture, the conditional probability of s is Gaussian:

$$\begin{aligned} p(s | \mathbf{x}, q_s, q_v) &= \frac{p(\mathbf{x}, s, q_s, q_v)}{p(\mathbf{x}, q_s, q_v)} \\ &= \frac{p(s|q_s)\, p(\mathbf{x}|s, q_v)\, p(q_s)\, p(q_v)}{p(\mathbf{x}|q_s, q_v)\, p(q_s)\, p(q_v)} \\ &= \frac{\mathbb{N}(s, \sigma_{s,q_s}^2)\ \mathbb{N}(\mathbf{x} - \mathbf{a}s, R_{v,q_v})}{\mathbb{N}\left(\mathbf{x}, R_{v,q_v} + \sigma_{s,q_s}^2 \mathbf{a}\mathbf{a}^H\right)} \\ &= \mathbb{N}\left(s - \alpha_{q_s,q_v},\ \beta_{q_s,q_v}\right) \end{aligned} \tag{34}$$

where

$$\alpha_{q_s,q_v} = \left(\sigma_{s,q_s}^{-2} + \mathbf{a}^H R_{v,q_v}^{-1}\mathbf{a}\right)^{-1}\mathbf{a}^H R_{v,q_v}^{-1}\mathbf{x} \quad (35)$$

$$\beta_{q_s,q_v} = \left(\sigma_{s,q_s}^{-2} + \mathbf{a}^H R_{v,q_v}^{-1}\mathbf{a}\right)^{-1} \quad (36)$$

M-step:: In the M-step, we maximize the expected complete likelihood with respect to the parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma_{s,q_s}^2, c_{v,q_v}, R_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$. This can be done by taking derivatives with respect to θ and setting them to be equal to zero (under the constraints $\sum_{q_s=1}^{k_s} c_{s,q_s} = 1$ and $\sum_{q_v=1}^{k_v} c_{v,q_v} = 1$). This results in the following update rules:

$$c_{v,q_v}^{(l)} = \frac{1}{\eta}\sum_{n=1}^{\eta}\sum_{q_s=1}^{k_s}\tau_{q_s,q_v}^{(l)}(n) \quad (37)$$

$$c_{s,q_s}^{(l)} = \frac{1}{\eta}\sum_{n=1}^{\eta}\sum_{q_v=1}^{k_v}\tau_{q_s,q_v}^{(l)}(n) \quad (38)$$

$$\sigma_{s,q_s}^{2^{(l)}} = \frac{\sum_{n=1}^{\eta}\sum_{q_v=1}^{k_v}\tau_{q_s,q_v}^{(l)}(n)\ \langle ss^*|\mathbf{x}(n), q_s, q_v\rangle}{\sum_{n=1}^{\eta}\sum_{q_v=1}^{k_v}\tau_{q_s,q_v}^{(l)}(n)} \quad (39)$$

$$R_{v,q_v}^{(l)} = \frac{\sum_{n=1}^{\eta}\sum_{q_s=1}^{k_s}\tau_{q_s,q_v}^{(l)}(n)\,\Lambda_{q_s,q_v}(n)}{\sum_{n=1}^{\eta}\sum_{q_s=1}^{k_s}\tau_{q_s,q_v}^{(l)}(n)} \quad (40)$$

where

$$\begin{aligned}\Lambda_{q_s,q_v}(n) &= \mathbf{x}(n)\mathbf{x}(n)^H - \mathbf{x}(n)\langle s^*|\mathbf{x}(n), q_s, q_v\rangle\,\mathbf{a}^H \\ &\quad -\mathbf{a}\langle s|\mathbf{x}(n), q_s, q_v\rangle\,\mathbf{x}(n)^H \\ &\quad +\mathbf{a}\langle ss^*|\mathbf{x}(n), q_s, q_v\rangle\,\mathbf{a}^H \end{aligned} \quad (41)$$

In this model, there is an ambiguity in associating variance between the desired source and the interference. It is possible to incorporate some of the source signal into the interference. To avoid this, updating the desired source component variances is not performed in the first few iterations. This prevents the source components shrinking to zero variance.

B. Learning Observed Mixture Parameters

In this section, the parameters $\theta_x = \{c_{x,q_x}, R_{x,q_x} : 1 \leq q_x \leq k_x)$ of the observed mixture $\mathbf{x}$ are estimated using the EM algorithm. These parameters are required for the non-linear beamformer $\mathbf{w}_3$ of

equation (22). Let us define a complete data set $D_c = \{\mathbf{x}, q_x\}$ composed of both the observed and the latent data. If we were to actually have such a complete data set, we define its likelihood as:

$$l_c(\theta_x|D_c) = \ln \prod_{n=1}^{\eta} p(\mathbf{x}(n), q_x(n)|\theta_x) = \sum_{n=1}^{\eta} \ln p(\mathbf{x}(n), q_x(n)|\theta_x) \tag{42}$$

The EM algorithm may be executed as follows:

E-step:: In the E-step, we compute the expectation of the complete data likelihood:

$$Q(\theta_x, \theta_x^{l-1}) = \sum_{q_x=1}^{k_x} p\left(q_x|\mathbf{x}, \theta_x^{l-1}\right) \ln p(\mathbf{x}, q_x|\theta_x) \tag{43}$$

This reduces to calculating $p\left(q_x|\mathbf{x}, \theta_x^{l-1}\right)$, the posterior probability of the latent variables given the observed data and the current estimates of the parameters.

$$\begin{aligned}
\tau_{q_x}^{(l)} &= p\left(q_x|\mathbf{x}, \theta_x^{(l-1)}\right) \\
&= \frac{p\left(q_x, \mathbf{x}|\theta_x^{(l-1)}\right)}{p\left(\mathbf{x}|\theta_x^{(l-1)}\right)} \\
&= \frac{p\left(q_x|\theta_x^{(l-1)}\right) p\left(\mathbf{x}|q_x, \theta_x^{(l-1)}\right)}{\sum_{q_x'=1}^{k_x} p\left(q_x'|\theta_x^{(l-1)}\right) p\left(\mathbf{x}|q_x', \theta_x^{(l-1)}\right)} \\
&= \frac{c_{q_x}^{(l-1)} \mathbb{N}\left(\mathbf{x}|R_{x,q_x}^{(l-1)}\right)}{\sum_{q_x'=1}^{k_x} c_{q_x'}^{(l-1)} \mathbb{N}\left(\mathbf{x}|R_{x,q_x'}^{(l-1)}\right)}
\end{aligned} \tag{44}$$

M-step:: In the M-step, we maximize the expected complete likelihood with respect to the parameters $\theta_x = \{c_{x,q_x}, R_{x,q_x} : 1 \leq q_x \leq k_x)$. This can be done by taking derivatives with respect to θ_x and setting them to be equal to zero, while also including a Lagrangian term to account for the constraint that $\sum_{q_x=1}^{k_x} c_{q_x} = 1$. This results in the following update rules:

$$R_{x,q_x}^{(l)} = \frac{\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n)\, \mathbf{x}(n)\, \mathbf{x}(n)^H}{\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n)} \tag{45}$$

$$c_{x,q_x}^{(l)} = \frac{1}{\eta} \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \tag{46}$$

ACKNOWLEDGMENT

The authors wish to thank S. Rickard for providing the implementation of the DUET algorithm.

REFERENCES

[1] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct 1992.

[2] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, ser. Signals and Communication Technology. Springer, 2005.

[3] S. Douglas and M. Gupta, "Convolutive blind source separation for audio signals," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer Netherlands, 2007.

[4] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer Netherlands, 2007.

[5] H. Attias, "Source separation with a sensor array using graphical models and subband filtering," in *Advances in Neural Inf. Process. Syst. (NIPS'02)*, 2002, pp. 1229–1236.

[6] ——, "New EM algorithms for source separation and deconvolution with a microphone array," in *IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'03)*, vol. 5, April 2003, pp. V–297–300 vol.5.

[7] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.

[8] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[9] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[10] M. Lockwood, D. Jones, R. Bilger, C. Lansing, W. O'Brien, Jr., B. Wheeler, and A. Feng, "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 379–391, 2004.

[11] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul 2004.

[12] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer Netherlands, 2007.

[13] Z. El-Chami, A. D. Pham, C. Servière, and A. Guerin, "A new model-based underdetermined speech separation," in *Proc. Intl. Workshop Acoust. Echo and Noise Control (IWAENC'08)*, Seattle, USA, Sep. 2008.

[14] D.-T. Pham, Z. El-Chami, A. Gurin, and C. Servire, "Modeling the short time Fourier transform ratio and application to underdetermined audio source separation." in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'09)*. Paraty, Brazil: Springer, 2009, pp. 98–105.

[15] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Inf. Process. Syst. (NIPS 19)*, B. Schlkopf, J. Platt, and T. Hoffman, Eds. MIT Press, 2006, pp. 953–960.

[16] M. Mandel and D. Ellis, "EM localization and separation using interaural level and phase cues," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'07)*, NY, USA, Oct. 2007, pp. 275–278.

[17] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'07)*, NY, USA, Oct. 2007, pp. 147–150.

[18] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. Intl. Workshop Acoust. Echo and Noise Control (IWAENC'05)*, Sep. 2005, pp. 117–120.

[19] ——, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[20] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[21] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 191–199, Jan. 2006.

[22] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'05)*, 2005, pp. 90–93.

[23] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1564–1578, Jul. 2007.

[24] H. Attias and L. Deng, "A new approach to speech enhancement by a microphone array using em and mixture models," in *Int.l Conf. on Spoken Lang. Process., Denver CO,*, 2002.

[25] H. L. Van Trees, *Optimum Array Processing*. John Wiley & Sons, Inc., 2002.

[26] M. A. Dmour and M. E. Davies, "An approach to under-determined speech separation based on a non-linear mixture of beamformers," in *Proc. Eur. Signal Process. Conf. (EUSIPCO'09)*, Glasgow, UK, Aug. 2009.

[27] H. Attias, "Independent factor analysis," *Neural Comput.*, vol. 11, no. 4, pp. 803–851, 1999.

[28] M. A. Dmour and M. E. Davies, "Under-determined speech separation using GMM-based non-linear beamforming," in *Proc. Eur. Signal Process. Conf. (EUSIPCO'08)*, Lausanne, Switzerland, Aug. 2008.

[29] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[30] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *DARPA Workshop on Speech Recognition*, 1986.

[31] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[32] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. Roy. Stat. Soc.*, vol. 46, no. 2, pp. 257–267, 1984.

[33] D.-S. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 27, no. 5, pp. 827–832, May 2005.

[34] K. Bell, Y. Ephraim, and H. Van Trees, "A Bayesian approach to robust adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 386–398, Feb. 2000.

[35] J. Burred and T. Sikora, "On the use of auditory representations for sparsity-based sound source separation," in *Proc. Int. Conf. Inform., Commun, Signal Process. (ICICS'05)*, Bangkok, Thailand, Dec. 2005, pp. 1466–1470.